(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号
特開2000－276308
（P2000－276308A）
(43)公開日　平成12年10月６日(2000.10.6)

(51)Int.Cl.⁷	識別記号	ＦＩ		テーマコード（参考）
Ｇ０６Ｆ　3/06	５４０	Ｇ０６Ｆ　3/06	５４０	５Ｂ０１８
	３０５		３０５Ａ	５Ｂ０６５
			３０５Ｃ	
			３０５Ｆ	
12/16	３１０	12/16	３１０Ｑ	

審査請求　未請求　請求項の数８　ＯＬ　（全 13 頁）　最終頁に続く

(21)出願番号　特願平11－85205

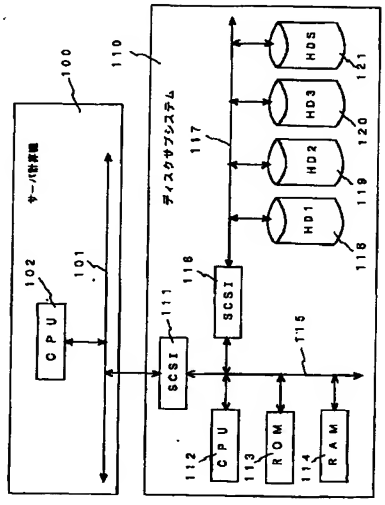
(22)出願日　平成11年３月29日(1999.3.29)

(71)出願人　000003078
株式会社東芝
神奈川県川崎市幸区堀川町72番地
(72)発明者　阿部　秀一
東京都青梅市末広町２丁目９番地　株式会社東芝青梅工場内
(74)代理人　100083161
弁理士　外川　英明

Ｆターム(参考)　5B018 GA02 GA06 HA04 HA13 HA14
HA35 MA12
5B065 BA01 CA30 EA02 EA12 EA24

(54)【発明の名称】　ディスクサブシステム及びディスクサブシステムにおけるデータ復元方法

(57)【要約】
【課題】本発明は、障害が発生したデータの復元処理に必要なデータが障害の発生していない磁気ディスク装置から読み出せないことに起因してディスクサブシステムが使用不能になることを防止するディスクサブシステムを提供することを目的とする。
【解決手段】ＲＡＩＤを構成したディスクサブシステムにおいて、障害の発生した磁気ディスク装置のデータを前記予備の磁気ディスク装置に復元する際に、前記障害の発生した磁気ディスク装置から読み出せるデータは、そのデータを前記予備の磁気ディスク装置に格納し、前記障害の発生した磁気ディスク装置から読み出せないデータは、前記ＲＡＩＤを構成する複数の磁気ディスク装置の内、障害が発生していない磁気ディスク装置に格納されているデータから演算により復元し、この復元したデータを前記予備の磁気ディスク装置に格納するデータ復元手段を設けたことを特徴とする。



【特許請求の範囲】
【請求項１】予備の磁気ディスク装置と複数の磁気ディスク装置とでＲＡＩＤを構成したディスクサブシステムにおいて、
障害の発生した磁気ディスク装置のデータを前記予備の磁気ディスク装置に復元する際に、前記障害の発生した磁気ディスク装置から読み出せるデータは、そのデータを前記予備の磁気ディスク装置に格納し、前記障害の発生した磁気ディスク装置から読み出せないデータは、前記ＲＡＩＤを構成する複数の磁気ディスク装置の内、障害が発生していない磁気ディスク装置に格納されているデータから演算により復元し、この復元したデータを前記予備の磁気ディスク装置に格納するデータ復元手段を具備したことを特徴とするディスクサブシステム。
【請求項２】前記データ復元手段は、演算によりデータを復元する際、その演算に用いられるデータが前記障害が発生していない磁気ディスク装置から読み出せない場合には、その復元対象のデータがＥＣＣエラーであると前記予備の磁気ディスク装置に登録すること特徴とする請求項１記載のディスクサブシステム。
【請求項３】予備の磁気ディスク装置と複数の磁気ディスク装置とでＲＡＩＤを構成したディスクサブシステムにおいて、
障害の発生した磁気ディスク装置のデータを前記予備の磁気ディスク装置に復元する際に、前記ＲＡＩＤを構成する複数の磁気ディスク装置の内、障害が発生していない磁気ディスク装置に格納されているデータから演算により復元し、この復元したデータを前記予備の磁気ディスク装置に格納するとともに、前記演算によりデータを復元するとき、その演算に用いられるデータが前記障害が発生していない磁気ディスク装置から読み出せない場合には、その復元対象のデータがＥＣＣエラーであると前記予備の磁気ディスク装置に登録するデータ復元手段を具備したことを特徴とするディスクサブシステム。
【請求項４】複数の磁気ディスク装置でＲＡＩＤが構成され、一つの磁気ディスク装置に障害が発生した場合にはその障害が発生した磁気ディスク装置を新しい磁気ディスク装置に交換するディスクサブシステムにおいて、前記障害の発生した磁気ディスク装置のデータを前記交換した新しい磁気ディスク装置に復元する際に、前記ＲＡＩＤを構成する複数の磁気ディスク装置の内、障害が発生していない磁気ディスク装置に格納されているデータから演算により復元し、この復元したデータを前記交換した新しい磁気ディスク装置に格納するとともに、前記演算によりデータを復元するとき、その演算に用いられるデータが前記障害が発生していない磁気ディスク装置から読み出せない場合には、その復元対象のデータがＥＣＣエラーであると前記交換した新しい磁気ディスク装置に登録するデータ復元手段を具備したことを特徴とするディスクサブシステム。
【請求項５】予備の磁気ディスク装置と複数の磁気ディスク装置とでＲＡＩＤを構成したディスクサブシステムにおいて、
障害の発生した磁気ディスク装置のデータを前記予備の磁気ディスク装置に復元する際に、前記障害の発生した磁気ディスク装置から読み出せるデータは、そのデータを前記予備の磁気ディスク装置に格納し、前記障害の発生した磁気ディスク装置から読み出せないデータは、前記ＲＡＩＤを構成する複数の磁気ディスク装置の内、障害が発生していない磁気ディスク装置に格納されているデータから演算により復元し、この復元したデータを前記予備の磁気ディスク装置に格納することを特徴とするディスクサブシステムにおけるデータ復元方法。
【請求項６】前記演算によりデータを復元するとき、その演算に用いられるデータが前記障害が発生していない磁気ディスク装置から読み出せない場合には、その復元対象のデータがＥＣＣエラーであると前記予備の磁気ディスク装置に登録すること特徴とする請求項５記載のディスクサブシステムにおけるデータ復元方法。
【請求項７】予備の磁気ディスク装置と複数の磁気ディスク装置とでＲＡＩＤを構成したディスクサブシステムにおいて、
障害の発生した磁気ディスク装置のデータを前記予備の磁気ディスク装置に復元する際に、前記ＲＡＩＤを構成する複数の磁気ディスク装置の内、障害が発生していない磁気ディスク装置に格納されているデータから演算により復元し、この復元したデータを前記予備の磁気ディスク装置に格納するとともに、前記演算によりデータを復元するとき、その演算に用いられるデータが前記障害が発生していない磁気ディスク装置から読み出せない場合には、その復元対象のデータがＥＣＣエラーであると前記予備の磁気ディスク装置に登録することを特徴とするディスクサブシステムにおけるデータ復元方法。
【請求項８】複数の磁気ディスク装置でＲＡＩＤが構成され、一つの磁気ディスク装置に障害が発生した場合にはその障害が発生した磁気ディスク装置を新しい磁気ディスク装置に交換するディスクサブシステムにおいて、前記障害の発生した磁気ディスク装置のデータを前記交換した新しい磁気ディスク装置に復元する際に、前記ＲＡＩＤを構成する複数の磁気ディスク装置の内、障害が発生していない磁気ディスク装置に格納されているデータから演算により復元し、この復元したデータを前記交換した新しい磁気ディスク装置に格納するとともに、前記演算によりデータを復元するとき、その演算に用いられるデータが前記障害が発生していない磁気ディスク装置から読み出せない場合には、その復元対象のデータがＥＣＣエラーであると前記交換した新しい磁気ディスク装置に登録することを特徴とするディスクサブシステムにおけるデータ復元方法。

【発明の詳細な説明】

【０００１】
【発明の属する技術分野】本発明は、ＰＣサーバ等のサーバ計算機において、高信頼性が要求される外部記憶装置として使用されるＲＡＩＤ技術を使用したディスクサブシステムの改良に関する。特に信頼性を更に高くするＲＡＩＤ技術を使用したディスクサブシステムに関する。
【０００２】
【従来の技術】ＲＡＩＤ（Ｒｅｄｕｎｄａｎｔ　Ａｒｒａｙｓ　ｏｆ　Ｉｎｅｘｐｅｎｓｉｖｅ　Ｄｉｓｋｓ）技術は、Ｄａｖｉｄ　Ａ．Ｐａｔｔｅｒｓｏｎ，ＧａｒｔｈＡ．Ｇｒｉｂｓｏｎ，Ｒａｎｄｙ　Ｈ．Ｋａｔｓにより１９８７年に発表された論文「Ａ　Ｃａｓｅ　ｆｏｒ　Ｒｅｄｕｎｄａｎｔ　Ａｒｒａｙｓ　ｏｆ　Ｉｎｅｘｐｅｎｓｉｖｅ　Ｄｉｓｋｓ　」で紹介されて実用化されている。
【０００３】そしてこのＲＡＩＤ技術は、一般的にレベル１からレベル５に分類され、高信頼性が要求されるＰＣサーバ等のサーバ計算機のディスクサブシステムに使用されている。
【０００４】レベル５のＲＡＩＤ技術を使用したディスクサブシステムの一例を図８を用いてその概要を説明する。図８において、ディスクサブシステムは、３台の磁気ディスク装置ＤＲ１、ＤＲ２、ＤＲ３と一台の予備磁気ディスク装置ＤＲＳとから構成されている。この図においては、サーバ計算機やＲＡＩＤコントローラの図示を省略している。
【０００５】サーバ計算機が管理し磁気ディスク装置に記録しているデータを所定単位のブロックに分割し、これを３台の磁気ディスク装置ＤＲ１、ＤＲ２、ＤＲ３に分散して記録する。この際に、冗長データとしてパリティを作成し、このパリティも分散して記録する。
【０００６】このパリティは、次のような方法で生成される。パリティＰ１は、データＤａとデータＤｂとの排他的論理和演算を行うことで生成される。同様にパリティＰ２はデータＤｃとデータＤｄとの排他的論理和演算で、パリティＰ３はデータＤｅとデータＤｆとの排他的論理和演算で、パリティＰ４はデータＤｇとデータＤｈとの排他的論理和演算で生成される。
【０００７】このように構成されたディスクサブシステムにおいて、サーバ計算機からデータＤｂのリード要求があった場合に、磁気ディスク装置ＤＲ２のデータＤｂが記録されているメディア（磁気ディスク）の記録面の障害又は磁気ディスク装置ＤＲ２全体の障害のため、データＤｂの読み出しができないときには、再度データリードを試みてリトライを実施し、それでも読み出しができない場合には、ＲＡＩＤコントローラがデータＤａとパリティＰ１とのの排他的論理和演算を行いデータＤｂを復元してサーバ計算機にリードデータとして出力してサーバ計算機にはディスクサブシステムに障害が発生したことを認識させずに対応している。
【０００８】このように障害が発生した場合、ＲＡＩＤコントローラは障害が発生した磁気ディスク装置ＤＲ２をＲＡＩＤ構成から切り離して磁気ディスク装置ＤＲ１とＤＲ３の２台の構成による縮退状態にする。
【０００９】この縮退状態で更に別な磁気ディスク装置に故障が発生すると、ディスクサブシステムからのデータの読み出しができなくなり、ＲＡＩＤコントローラはサーバ計算機に対してディスクサブシステムに障害が発生したと通知することになる。
【００１０】このように縮退状態で更に別な磁気ディスク装置に故障が発生してディスクサブシステムに障害が発生させないように、ＲＡＩＤを正常な状態に回復する必要がある。
【００１１】この回復処理は、上記縮退状態において予備の磁気ディスク装置ＤＲＳに切り離した磁気ディスク装置ＤＲ２のデータを復元することで行われる。この予備の磁気ディスク装置ＤＲＳに切り離した磁気ディスク装置ＤＲ２のデータを復元するには、磁気ディスク装置ＤＲ１とＤＲ３のそれぞれに対応するブロックのデータどうしの排他的論理和演算を行うことで実現できる。例えばデータＤｅを復元するには、パリティＰ３とデータＤｆとの排他的論理和演算で復元できる。
【００１２】また、予備の磁気ディスク装置を持たないディスクサブシステムの回復処理では、障害の発生した磁気ディスク装置を別の正常な磁気ディスク装置と交換し、この交換した磁気ディスク装置にデータを上記と同様に排他的論理和演算により復元する。
【００１３】
【発明が解決しようとする課題】このようなＲＡＩＤ技術を使用したディスクサブシステムにおいては、次のような問題点があった。障害が発生しＲＡＩＤ構成から切り離した磁気ディスク装置のデータを予備の又は交換した磁気ディスク装置に復元する際に、上記切り離した磁気ディスク装置の障害の内容がメディア（磁気ディスク）の記録面の場合には、障害の発生していないメディアからはそこに記録されているデータの読み出しが可能にもかかわらず、読み出しが可能なデータも含めて全てのデータを排他的論理和演算により求めるため、データの復元処理に大幅な時間を要していた。
【００１４】また、このようなデータの復元処理の際に、排他的論理和演算のためにＲＡＩＤ構成している障害の発生していない磁気ディスク装置からデータを読み出す際に、障害が発生してデータの読み出しができない場合がある。このとき、この時点でデータの復元処理は不可能になるばかりでなく、ディスクサブシステム自体が使用不可になってしまっていた。このようにデータの復元処理の際に障害が見つかるデータは、普段アクセスがされていないあまり重要なデータではないと推測される。このようなあまり重要でないデータが読み出せない

ことに起因して、障害の発生した磁気ディスク装置のデータの復元ができなくなるとともに、ディスクサブシステム自体が使用不能になることは大きな問題であった。

【００１５】本発明は、これら従来の問題点を解決するためになされたもので、障害が発生した磁気ディスク装置のデータの復元処理を高速化できるディスクサブシステムを提供することを目的とする。

【００１６】また、本発明は、障害が発生した磁気ディスク装置のデータの復元処理の際に、データの復元処理に必要なデータが障害の発生していない磁気ディスク装置から読み出せないことに起因してＲＡＩＤが再構成できなくなり、ディスクサブシステムが使用不能になることを防止しするディスクサブシステムを提供することを目的とする。

【００１７】

【課題を解決するための手段】本発明は、予備の磁気ディスク装置と複数の磁気ディスク装置とでＲＡＩＤを構成したディスクサブシステムにおいて、障害の発生した磁気ディスク装置のデータを前記予備の磁気ディスク装置に復元する際に、前記障害の発生した磁気ディスク装置から読み出せるデータは、そのデータを前記予備の磁気ディスク装置に格納し、前記障害の発生した磁気ディスク装置から読み出せないデータは、前記ＲＡＩＤを構成する複数の磁気ディスク装置の内、障害が発生していない磁気ディスク装置に格納されているデータから演算により復元し、この復元したデータを前記予備の磁気ディスク装置に格納するデータ復元手段を設けたことを特徴とする。

【００１８】このような構成によれば、データの復元処理を高速化できる。また、本発明は、予備の磁気ディスク装置と複数の磁気ディスク装置とでＲＡＩＤを構成したディスクサブシステムにおいて、障害の発生した磁気ディスク装置のデータを前記予備の磁気ディスク装置に復元する際に、前記障害の発生した磁気ディスク装置から読み出せるデータは、そのデータを前記予備の磁気ディスク装置に格納し、前記障害の発生した磁気ディスク装置から読み出せないデータは、前記ＲＡＩＤを構成する複数の磁気ディスク装置の内、障害が発生していない磁気ディスク装置に格納されているデータから演算により復元し、この復元したデータを前記予備の磁気ディスク装置に格納するデータ復元手段を設け、更に前記演算によりデータを復元する際、その演算に用いられるデータが前記障害が発生していない磁気ディスク装置から読み出せない場合には、その復元対象のデータがＥＣＣエラーであると前記予備の磁気ディスク装置に登録することを特徴とする。

【００１９】このような構成によれば、データの復元処理を高速化できるとともに障害が発生した磁気ディスク装置のデータの復元処理の際に、データの復元処理に必要なデータが障害の発生していない磁気ディスク装置から読み出せないことに起因してＲＡＩＤが再構成できなくなり、ディスクサブシステムが使用不能になることを防止できる。

【００２０】また、本発明は、予備の磁気ディスク装置と複数の磁気ディスク装置とでＲＡＩＤを構成したディスクサブシステムにおいて、障害の発生した磁気ディスク装置のデータを前記予備の磁気ディスク装置に復元する際に、前記ＲＡＩＤを構成する複数の磁気ディスク装置の内、障害が発生していない磁気ディスク装置に格納されているデータから演算により復元し、この復元したデータを前記予備の磁気ディスク装置に格納するとともに、前記演算によりデータを復元するとき、その演算に用いられるデータが前記障害が発生していない磁気ディスク装置から読み出せない場合には、その復元対象のデータがＥＣＣエラーであると前記予備の磁気ディスク装置に登録するデータ復元手段を設けたことを特徴とする。

【００２１】このような構成によれば、障害が発生した磁気ディスク装置のデータの復元処理の際に、データの復元処理に必要なデータが障害の発生していない磁気ディスク装置から読み出せないことに起因してＲＡＩＤが再構成できなくなり、ディスクサブシステムが使用不能になることを防止できる。

【００２２】更に、本発明は、複数の磁気ディスク装置でＲＡＩＤが構成され、一つの磁気ディスク装置に障害が発生した場合にはその障害が発生した磁気ディスク装置を新しい磁気ディスク装置に交換するディスクサブシステムにおいて、前記障害の発生した磁気ディスク装置のデータを前記交換した新しい磁気ディスク装置に復元する際に、前記ＲＡＩＤを構成する複数の磁気ディスク装置の内、障害が発生していない磁気ディスク装置に格納されているデータから演算により復元し、この復元したデータを前記交換した新しい磁気ディスク装置に格納するとともに、前記演算によりデータを復元するとき、その演算に用いられるデータが前記障害が発生していない磁気ディスク装置から読み出せない場合には、その復元対象のデータがＥＣＣエラーであると前記交換した新しい磁気ディスク装置に登録するデータ復元手段を設けたことを特徴とする。

【００２３】このような構成によれば、障害が発生した磁気ディスク装置のデータの復元処理の際に、データの復元処理に必要なデータが障害の発生していない磁気ディスク装置から読み出せないことに起因してＲＡＩＤが再構成できなくなり、ディスクサブシステムが使用不能になることを防止できる。

【００２４】

【発明の実施の形態】以下、図面を参照して本発明の第１の実施形態について説明する。図１は本発明の第１の実施形態に関わるシステムの概略構成を説明する図である。図１には、ＰＣサーバなどのサーバ計算機１００と

ディスクサブシステム１１０とからなる計算機システムの概略構成を図示している。

【００２５】サーバ計算機１００にはシステムバス１０１が設けられており、このシステムバス１０１にＣＰＵ１０２が接続されている。また、システムバス１０１にはＳＣＳＩインタフェース（Ｉ／Ｆ）１１１を介してディスクサブシステム１１０が接続されている。尚、ＳＣＳＩは、Ｓｍａｌｌ　Ｃｏｍｐｕｔｅｒ　Ｓｙｓｔｅｍ　Ｉｎｔｅｒｆａｃｅの省略語である。

【００２６】ディスクサブシステム１１０は、ＣＰＵ１１２、ＲＯＭ１１３、ＲＡＭ１１４、ＳＣＳＩインタフェース（Ｉ／Ｆ）１１１とＳＣＳＩインタフェース（Ｉ／Ｆ）１１６がバス１１５を介して相互に接続されているＲＡＩＤ制御部とＳＣＳＩバス１１７を介してＳＣＳＩインタフェース（Ｉ／Ｆ）１１６に接続された４台の磁気ディスク装置ＨＤ１、ＨＤ２、ＨＤ３、ＨＤＳとから構成されている。

【００２７】ＣＰＵ１１２は、ＲＯＭ１１３に格納されたファームウェアに基づいて、ディスクサブシステム１１０の全体を制御する。ＲＡＭ１１４には、サーバ計算機１００からディスクサブシステム１１０に対するデータのリード又はライトの命令が出された際のアドレス情報を実際の磁気ディスク装置に記録されているアドレス情報に変換する変換テーブルが設けられている。また、ＲＡＭ１１４は、ＣＰＵ１１２の動作に伴う各種データが保存される。磁気ディスク装置ＨＤ１と磁気ディスク装置ＨＤ２と磁気ディスク装置ＨＤ３とでレベル５のＲＡＩＤを構成している。そして、磁気ディスク装置ＨＤＳは、予備の磁気ディスク装置として設けられている。

【００２８】図２には、サーバ計算機１００から見える論理的にディスクサブシステム１１０に記録されているデータの配置の構成と実際にＲＡＩＤを構成する個々の磁気ディスク装置に記録されているデータとの関係を図示している。

【００２９】サーバ計算機１００から見て論理的にディスクサブシステム１１０に記録されているデータ２００は、所定サイズの６つのブロックに分割されたデータＤ１～Ｄ６に分割されている。この６つのデータＤ１～Ｄ６は、図示のように３台の磁気ディスク装置ＨＤ１、ＨＤ２、ＨＤ３に記録されている。磁気ディスク装置ＨＤ１には、パリティデータＰ１、データＤ３、データＤ５とが記録されている。同様に、磁気ディスク装置ＨＤ２には、データＤ２、パリティデータＰ２、データＤ６とが記録されている。また、磁気ディスク装置ＨＤ３には、データＤ２、データＤ４、パリティデータＰ３とが記録されている。磁気ディスク装置ＨＤＳは、予備の磁気ディスク装置であるため、現時点では何も記録されていない。パリティＰ１は、データＤ１とデータＤ２との排他的論理和演算により求められたデータである。同様にパリティＰ２は、データＤ３とデータＤ４との排他的論理和演算により求められたデータである。また、パリティＰ３は、データＤ５とデータＤ６との排他的論理和演算により求められたデータである。

【００３０】以下、図３、図４、図５に示すフローチャートに基づいて、本発明の第１の実施形態におけるデータの復元動作について説明する。尚、データの復元動作中においても、並行してサーバ計算機１００からのディスクサブシステム１１０へのデータの書き込み及び読み出しの命令を受け付け、その処理を行うものとする。復元動作中にサーバ計算機１００からデータの書き込み又は読み出しの命令を受け付けた場合の処理は、以下の動作説明の中で説明する。

【００３１】以下、レベル５のＲＡＩＤを構成している図１に図示したディスクサブシステム１１０において、メディアの障害により磁気ディスク装置ＨＤ３からデータＤ２の読み出しができなくなり、ＲＡＩＤ制御装置のＣＰＵ１１２が磁気ディスク装置ＨＤ３をＲＡＩＤ構成から切り離して縮退状態とし、予備の磁気ディスク装置ＨＤＳに磁気ディスク装置ＨＤ３に格納されていたデータを復元する場合のＣＰＵ１１２の制御に基づく動作を説明する。

【００３２】まず、磁気ディスク装置ＨＤ３に格納されていたデータを予備の磁気ディスク装置ＨＤＳに復元するに際して、ＣＰＵ１１２はＲＡＭ１１４上に図６に示すようなデータ変更／復元マップ３００を作成する（ステップＳ１）。このデータ変更／復元マップ３００は、データ３０１、状態ビット３０２、復元ビット３０３の３行から構成されており、データを復元する磁気ディスク装置ＨＤ３の復元動作中におけるデータの変更及び復元の状態を管理するために設けられている。従って、データ３０１としては、Ｄ２、Ｄ４、Ｐ３が登録される。また、この初期時には状態ビット３０２と復元ビット３０３は、「０」に設定される。復元ビット３０３は、以下に説明するデータの復元処理が終了した状態をデータ（ブロック）毎に表すものである。このビットが「１」に設定されているということは、その対応するデータが予備の磁気ディスク装置ＨＤＳに復元されたことを意味する。また、状態ビット３０２は、この復元動作中にサーバ計算機１００からの書き込み命令により、対応するブロックのデータが変更された場合にセットされる。

【００３３】次に変数Ｂを「１」を設定する（ステップＳ２）。この変数Ｂは、磁気ディスク装置ＨＤ３の何番目のブロックのデータを復元動作するかを示すものである。次にフラグ情報である変数ＳＥＴを「１」を設定する（ステップＳ３）。この変数ＳＥＴが「１」にセットされている場合は、ＣＰＵ１１２が磁気ディスク装置ＨＤ３のブロック単位のデータを予備の磁気ディスク装置ＨＤＳに復元動作中であることを示す。次にＢ番目のブロックのデータの復元処理を行う（ステップＳ４）。即ち、１番目のブロック（データＤ１が格納されているブ

ロック）のデータ（データＤ１）の復元処理を行う。
【００３４】この復元処理の詳細は、図７に示すフローチャートを用いて説明する。まず、該当する１番目のブロックの状態ビットに「１」が設定されているかどうかをデータ変更／復元マップ３００を参照してチェックする（ステップＳ５０）。このチェックの結果、状態ビットに「１」が設定されている場合には、ステップＳ５２へ進む。現時点では、状態ビットに「０」が設定されているため、ステップＳ５１に進む。
【００３５】次に、１番目のブロックのデータＤ２が障害の発生した磁気ディスク装置ＨＤ３から読み出せるかをチェックする（ステップＳ５１）。読み出せる場合には、磁気ディスク装置ＨＤ３からデータＤ２を読み出し、予備の磁気ディスク装置ＨＤＳの対応するブロックにコピーして復元する（ステップＳ５３）。そして、復元動作を終了する。もし、読み出せない場合には、ステップＳ５２に進む。
【００３６】次に１番目のブロックのデータＤ２が排他的論理和演算で復元できるかをチェックする。換言すると、この排他的論理和演算をするためのデータであるパリティＰ１とデータＤ１とが、それぞれ磁気ディスク装置ＨＤ１及び磁気ディスク装置ＨＤ２から読み出せるかをチェックする（ステップＳ５２）。読み出せる場合には、ステップＳ５４へ進み、パリティＰ１とデータＤ１との排他的論理和演算を行い、データＤ２を復元し、予備の磁気ディスク装置ＨＤＳの対応するブロックにコピーする（ステップＳ５４）。そして、復元動作を終了する。また、読み出せない場合には、ステップＳ５５へ進む。
【００３７】ステップＳ５５では、現在復元処理をしているブロックのデータが、障害の発生した磁気ディスク装置ＨＤ３から読み出せず、更に排他的論理和演算によっても復元できない状態のため、予備の磁気ディスク装置ＨＤＳの対応するブロックにＥＣＣ（Ｅｒｒｏｒ　Ｃｈｅｃｋｉｎｇ　ａｎｄ　Ｃｏｒｒｅｃｔｉｎｇ）エラーが発生したとして、予備の磁気ディスク装置ＨＤＳのメディアの特定エリアに設定した図示しないエラー状態マップに該当するブロックにＥＣＣエラーが発生したことのフラグを登録する。この様な一連の動作でデータの復元処理を行い、図３のステップＳ４のデータの復元処理が終了する。
【００３８】次に図３のステップＳ５に進む。ステップＳ５では、ステップＳ４においてデータＤ２が復元されたので、図６に示すデータＤ２に対応する復元ビットを「１」にセットする。
【００３９】次にステップＳ６では、変数Ｂを＋１してインクリメントする。続いてステップＳ７において、変数ＳＥＴを「０」を設定する。この変数ＳＥＴを「０」にセットすることで、、ＣＰＵ１１２が磁気ディスク装置ＨＤ３のブロック単位のデータを予備の磁気ディスク装置ＨＤＳに復元する動作が終了したことを意味する。
【００４０】続いて処理はステップＳ８に進み、既に復元したブロックのデータがサーバ計算機１００からのデータの書き込み命令により変更されていないかをチェックする。具体的には、図６に示したデータ変更／復元マップ３００において、復元ビット３０３が「１」に設定されているデータの内、状態ビット３０２が「１」に設定されているデータの有無をチェックする。このチェックの結果、復元ビット３０３と状態ビット３０２の双方が「１」に設定されているデータがある場合には、そのデータが既にデータが復元されたもので、且つその後のサーバ計算機１００からのデータの書き込み命令により変更されているものであると判断されステップＳ１０へ進む。また、復元ビット３０３と状態ビット３０２の双方が「１」に設定されているデータがない場合には、ステップＳ９に進む。以下、ステップＳ９に進む場合の動作を説明し、ステップＳ１０へ進む場合の動作は、後に説明する。
【００４１】ステップＳ９では、全てのブロックのデータが復元されたかを変数Ｂの値を見ることでチェックする。全てのブロックのデータが復元された場合には、全ての復元処理が終了する。全てのブロックのデータが復元されていない場合には、ステップＳ３へ戻り以上説明した動作を繰り返す。
【００４２】次に以上説明したステップＳ３からＳ９の処理を繰り返している間にサーバ計算機１００からディスクサブシステム１１０へのデータの書き込み又は読み出しの命令をＣＰＵ１１２がＳＣＳＩインタフェース１１１を介して受理した場合には、以上説明したデータの復元処理動作を中断して、そのデータの書き込み又は読み出し処理を実行し、その後復元処理動作を再開する。
【００４３】まず、ＣＰＵ１１２がサーバ計算機１００からデータの読み出し命令を受理した場合について説明する。ＲＡＩＤを構成する正常な磁気ディスク装置からのデータの読み出しであれば、そのデータを読み出す。もし、障害が発生しＲＡＩＤ構成から切り離した磁気ディスク装置ＨＤ３からのデータの読み出しの場合には、もしそのデータが磁気ディスク装置ＨＤ３から読み出せる場合には、そのまま読み出し、読み出せない場合には、他の正常な磁気ディスク装置から読み出した対応するブロックのデータの排他的論理和演算により求めたデータを読み出しデータとしてサーバ計算機１００からへ転送する。ただし、このように説明したいずれの読み出しの場合においても、メディアの障害が原因によりにデータが読み出せないときには、上述したようにＥＣＣエラーとして登録し、サーバ計算機１００からにその旨を通知する。
【００４４】次にＣＰＵ１１２がサーバ計算機１００からデータの書き込み命令を受理した場合について説明する。まず、上記変数ＳＥＴが「０」であることを確認す

る。もし、上記変数ＳＥＴが「１」の場合には、「０」になるまで、その処理を待機させる。変数ＳＥＴが「０」であることを確認できた場合には、以下の処理を行う。ＲＡＩＤを構成する正常な磁気ディスク装置に記録されているデータの書き込みであれば、そのまま書き込み処理を実施する。この際、パリティの変更をすることは説明するまでもない。

【００４５】また、障害が発生しＲＡＩＤ構成から切り離した磁気ディスク装置ＨＤ３へのデータの書き込みの場合、例えばデータＤ４の書き込みの場合には、次のように行う。まず、この新しい書き込みデータＤ４と磁気ディスク装置ＨＤ１に書き込まれているデータＤ３とで排他的論理和演算を行い得られたパリティデータを新しいパリティＰ２として磁気ディスク装置ＨＤ２に書き込む。そして、新しい書き込みデータＤ４を実際に書き込む代わりに図６に示したデータ変更／復元マップ３００のデータＤ４に対応する状態ビット３０２を「１」にセットする。この様な処理をすることで書き込み処理を実施する。

【００４６】次に図４のステップＳ８において、復元ビット３０３と状態ビット３０２の双方が「１」に設定されており、既にそのデータが復元されたもので、且つその後サーバ計算機１００からのデータの書き込み命令により変更されているデータがあると判断されステップＳ１０へ進む場合の動作を説明する。以下の説明ではデータ変更／復元マップ３００が図４に示した状態になっている場合を例にして説明する。

【００４７】ステップＳ１０では、フラグ情報である変数ＳＥＴを「１」を設定する。続いて、ステップＳ１１では、ステップＳ８において復元ビット３０３と状態ビット３０２の双方が「１」に設定されていると判断されたデータＤ２の復元処理を再度実行する。この復元処理は、図７で示したルーチンに基づき行われるもので、既に説明したので、ここではその説明を省略する。

【００４８】次にステップＳ１２で図６に示したデータ変更／復元マップ３００におけるデータＤ２に対応する状態ビット３０２を「０」に設定する。続いてステップＳ１３において、変数ＳＥＴを「０」を設定しステップＳ８に戻る。この様に復元ビット３０３と状態ビット３０２の双方が「１」に設定されているデータが無くなるまで、ステップＳ８→Ｓ１０→Ｓ１１→Ｓ１２→Ｓ１３の処理を続ける。

【００４９】次に本発明の第２の実施形態について説明する。第１の実施形態との違いは、図７に示した復元処理の動作が異なり、他の動作は、第１の実施形態と同一である。

【００５０】第２の実施形態における復元処理の動作を図８に示したフローチャートを用いて説明する。動作説明の前提としては、第１の実施形態と同様にレベル５のＲＡＩＤを構成している図１に図示したディスクサブシステム１１０において、メディアの障害により磁気ディスク装置ＨＤ３からデータＤ２の読み出しができなくなり、ＲＡＩＤ制御装置のＣＰＵ１１２が磁気ディスク装置ＨＤ３をＲＡＩＤ構成から切り離して縮退状態とし、予備の磁気ディスク装置ＨＤＳに磁気ディスク装置ＨＤ３に格納されていたデータを復元する場合のＣＰＵ１１２の制御に基づく動作を説明する。

【００５１】まず、１番目のブロックのデータＤ２が排他的論理和演算で復元できるかをチェックする。換言すると、この排他的論理和演算をするためのデータであるパリティＰ１とデータＤ１とが、それぞれ磁気ディスク装置ＨＤ１及び磁気ディスク装置ＨＤ２から読み出せるかをチェックする（ステップＳ６０）。読み出せる場合には、ステップＳ６１へ進み、パリティＰ１とデータＤ１との排他的論理和演算を行い、データＤ２を復元し、予備の磁気ディスク装置ＨＤＳの対応するブロックにコピーする（ステップＳ６１）。そして、復元動作を終了する。

【００５２】また、読み出せない場合には、ステップＳ６２へ進む。ステップＳ６２では、１番目のブロックのデータＤ２が障害の発生した磁気ディスク装置ＨＤ３から読み出せるかをチェックする（ステップＳ６２）。読み出せる場合には、磁気ディスク装置ＨＤ３からデータＤ２を読み出し、予備の磁気ディスク装置ＨＤＳの対応するブロックにコピーして復元する（ステップＳ６３）。そして、復元動作を終了する。

【００５３】もし、読み出せない場合には、ステップＳ６４に進む。ステップＳ６４では、現在復元処理をしているブロックのデータが、排他的論理和演算によっても復元できず、更に障害の発生した磁気ディスク装置ＨＤ３から読み出せない状態のため、予備の磁気ディスク装置ＨＤＳの対応するブロックにＥＣＣエラーが発生したとして、予備の磁気ディスク装置ＨＤＳのメディアの特定エリアに設定した図示しないエラー状態マップに該当するブロックにＥＣＣエラーが発生したことのフラグを登録する。この様な一連の動作でデータの復元処理を行い、図３のステップＳ４のデータの復元処理が終了する。

【００５４】次に本発明の第３の実施形態について説明をする。第１の実施形態との違いは、図７に示した復元処理の動作が異なることと、ディスクサブシステム１１０に予備の磁気ディスク装置を持たず、ＲＡＩＤを構成する磁気ディスク装置に障害が発生した場合には、その磁気ディスク装置を新しい正常な磁気ディスク装置と交換することである。その他の動作は、第１の実施形態と同一である。

【００５５】第３の実施形態における復元処理の動作を図９に示したフローチャートを用いて説明する。動作説明の前提としては、第１の実施形態と同様にレベル５のＲＡＩＤを構成している図１に図示したディスクサブシ

る。もし、上記変数ＳＥＴが「１」の場合には、「０」になるまで、その処理を待機させる。変数ＳＥＴが「０」であることを確認できた場合には、以下の処理を行う。ＲＡＩＤを構成する正常な磁気ディスク装置に記録されているデータの書き込みであれば、そのまま書き込み処理を実施する。この際、パリティの変更をすることは説明するまでもない。

【００４５】また、障害が発生しＲＡＩＤ構成から切り離した磁気ディスク装置ＨＤ３へのデータの書き込みの場合、例えばデータＤ４の書き込みの場合には、次のように行う。まず、この新しい書き込みデータＤ４と磁気ディスク装置ＨＤ１に書き込まれているデータＤ３とで排他的論理和演算を行い得られたパリティデータを新しいパリティＰ２として磁気ディスク装置ＨＤ２に書き込む。そして、新しい書き込みデータＤ４を実際に書き込む代わりに図６に示したデータ変更／復元マップ３００のデータＤ４に対応する状態ビット３０２を「１」にセットする。この様な処理をすることで書き込み処理を実施する。

【００４６】次に図４のステップＳ８において、復元ビット３０３と状態ビット３０２の双方が「１」に設定されており、既にそのデータが復元されたもので、且つその後サーバ計算機１００からのデータの書き込み命令により変更されているデータがあると判断されステップＳ１０へ進む場合の動作を説明する。以下の説明ではデータ変更／復元マップ３００が図４に示した状態になっている場合を例にして説明する。

【００４７】ステップＳ１０では、フラグ情報である変数ＳＥＴを「１」を設定する。続いて、ステップＳ１１では、ステップＳ８において復元ビット３０３と状態ビット３０２の双方が「１」に設定されていると判断されたデータＤ２の復元処理を再度実行する。この復元処理は、図７で示したルーチンに基づき行われるもので、既に説明したので、ここではその説明を省略する。

【００４８】次にステップＳ１２で図６に示したデータ変更／復元マップ３００におけるデータＤ２に対応する状態ビット３０２を「０」に設定する。続いてステップＳ１３において、変数ＳＥＴを「０」を設定しステップＳ８に戻る。この様に復元ビット３０３と状態ビット３０２の双方が「１」に設定されているデータが無くなるまで、ステップＳ８→Ｓ１０→Ｓ１１→Ｓ１２→Ｓ１３の処理を続ける。

【００４９】次に本発明の第２の実施形態について説明する。第１の実施形態との違いは、図７に示した復元処理の動作が異なり、他の動作は、第１の実施形態と同一である。

【００５０】第２の実施形態における復元処理の動作を図８に示したフローチャートを用いて説明する。動作説明の前提としては、第１の実施形態と同様にレベル５のＲＡＩＤを構成している図１に図示したディスクサブシステム１１０において、メディアの障害により磁気ディスク装置ＨＤ３からデータＤ２の読み出しができなくなり、ＲＡＩＤ制御装置のＣＰＵ１１２が磁気ディスク装置ＨＤ３をＲＡＩＤ構成から切り離して縮退状態とし、予備の磁気ディスク装置ＨＤＳに磁気ディスク装置ＨＤ３に格納されていたデータを復元する場合のＣＰＵ１１２の制御に基づく動作を説明する。

【００５１】まず、１番目のブロックのデータＤ２が排他的論理和演算で復元できるかをチェックする。換言すると、この排他的論理和演算をするためのデータであるパリティＰ１とデータＤ１とが、それぞれ磁気ディスク装置ＨＤ１及び磁気ディスク装置ＨＤ２から読み出せるかをチェックする（ステップＳ６０）。読み出せる場合には、ステップＳ６１へ進み、パリティＰ１とデータＤ１との排他的論理和演算を行い、データＤ２を復元し、予備の磁気ディスク装置ＨＤＳの対応するブロックにコピーする（ステップＳ６１）。そして、復元動作を終了する。

【００５２】また、読み出せない場合には、ステップＳ６２へ進む。ステップＳ６２では、１番目のブロックのデータＤ２が障害の発生した磁気ディスク装置ＨＤ３から読み出せるかをチェックする（ステップＳ６２）。読み出せる場合には、磁気ディスク装置ＨＤ３からデータＤ２を読み出し、予備の磁気ディスク装置ＨＤＳの対応するブロックにコピーして復元する（ステップＳ６３）。そして、復元動作を終了する。

【００５３】もし、読み出せない場合には、ステップＳ６４に進む。ステップＳ６４では、現在復元処理をしているブロックのデータが、排他的論理和演算によっても復元できず、更に障害の発生した磁気ディスク装置ＨＤ３から読み出せない状態のため、予備の磁気ディスク装置ＨＤＳの対応するブロックにＥＣＣエラーが発生したとして、予備の磁気ディスク装置ＨＤＳのメディアの特定エリアに設定した図示しないエラー状態マップに該当するブロックにＥＣＣエラーが発生したことのフラグを登録する。この様な一連の動作でデータの復元処理を行い、図３のステップＳ４のデータの復元処理が終了する。

【００５４】次に本発明の第３の実施形態について説明をする。第１の実施形態との違いは、図７に示した復元処理の動作が異なることと、ディスクサブシステム１１０に予備の磁気ディスク装置を持たず、ＲＡＩＤを構成する磁気ディスク装置に障害が発生した場合には、その磁気ディスク装置を新しい正常な磁気ディスク装置と交換することである。その他の動作は、第１の実施形態と同一である。

【００５５】第３の実施形態における復元処理の動作を図９に示したフローチャートを用いて説明する。動作説明の前提としては、第１の実施形態と同様にレベル５のＲＡＩＤを構成している図１に図示したディスクサブシ

ステム１１０において、メディアの障害により磁気ディスク装置ＨＤ３からデータＤ２の読み出しができなくなり、磁気ディスク装置ＨＤ３を新しい正常な磁気ディスク装置と交換し、この交換した新しい磁気ディスク装置（図示せず）に磁気ディスク装置ＨＤ３に格納されていたデータを復元する場合のＣＰＵ１１２の制御に基づく動作を説明する。

【００５６】まず、１番目のブロックのデータＤ２が排他的論理和演算で復元できるかをチェックする。換言すると、この排他的論理和演算をするためのデータであるパリティＰ１とデータＤ１とが、それぞれ磁気ディスク装置ＨＤ１及び磁気ディスク装置ＨＤ２から読み出せるかをチェックする（ステップＳ７０）。読み出せる場合には、ステップＳ７１へ進み、パリティＰ１とデータＤ１との排他的論理和演算を行い、データＤ２を復元し、交換した新しい磁気ディスク装置の対応するブロックにコピーする（ステップＳ７１）。そして、復元動作を終了する。

【００５７】また、読み出せない場合には、ステップＳ７２へ進む。ステップＳ７２では、現在復元処理をしているブロックのデータが、排他的論理和演算によっても復元できないため、交換した新しい磁気ディスク装置の対応するブロックにＥＣＣエラーが発生したとして、交換した新しい磁気ディスク装置のメディアの特定エリアに設定した図示しないエラー状態マップに該当するブロックにＥＣＣエラーが発生したことのフラグを登録する。この様な一連の動作でデータの復元処理を行い、図３のステップＳ４のデータの復元処理が終了する。

【００５８】以上の説明では、復元処理をしているブロックのデータが、復元できないため、ＥＣＣエラーが発生したとして、データの復元をしている磁気ディスク装置にＥＣＣエラーが発生したことのフラグを登録すると説明した。しかし、その後、その復元できなかったデータがサーバ計算機１００により新たにデータが書き込まれた場合には、そのフラグの登録を削除すればよい。

【００５９】尚、このようにＥＣＣエラーが発生したことのフラグが登録されるデータは、従来技術の説明で述べたように、普段アクセスがされていないあまり重要なデータではないと推測される。従って、データが復元できなくてＲＡＩＤの再構成ができなくなり、ディスクサブシステム自体が使用不能になることに比べて、復元できないデータにＥＣＣエラーが発生したと登録する処理の方が、システム全体から見れば、より良い対策である。

【００６０】

【発明の効果】以上説明した通り、本発明によれば、データの復元処理を高速化できるとともに、障害が発生した磁気ディスク装置のデータの復元処理の際に、データの復元処理に必用なデータが障害の発生していない磁気ディスク装置から読み出せないことに起因してＲＡＩＤが再構成できなくなり、ディスクサブシステムが使用不能になることを防止できる。

【図面の簡単な説明】

【図１】本発明の第１の実施形態に関わるシステムの概略構成を示す図である。

【図２】サーバ計算機から見える論理的にディスクサブシステムに記録されているデータの配置の構成と実際にＲＡＩＤを構成する個々の磁気ディスク装置に記録されているデータとの関係を示す図。

【図３】本発明の第１の実施形態におけるデータの復元動作を説明するためのフローチャート図。

【図４】本発明の第１の実施形態におけるデータの復元動作を説明するためのフローチャート図。

【図５】本発明の第１の実施形態におけるデータの復元動作を説明するためのフローチャート図。

【図６】データ変更／復元マップを示す図。

【図７】図３におけるデータ復元処理の詳細を説明するためのフローチャート図。

【図８】本発明の第２の実施形態におけるデータ復元処理の詳細を説明するためのフローチャート図。

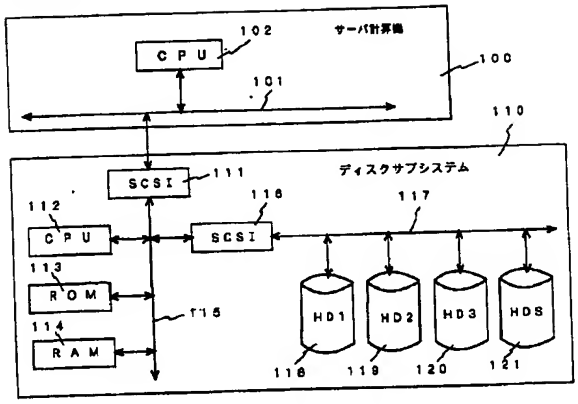
【図９】本発明の第３の実施形態におけるデータ復元処理の詳細を説明するためのフローチャート図。

【図１０】従来技術を説明するための、サーバ計算機から見える論理的にディスクサブシステムに記録されているデータの配置の構成と実際にＲＡＩＤを構成する個々の磁気ディスク装置に記録されているデータとの関係を示す図。

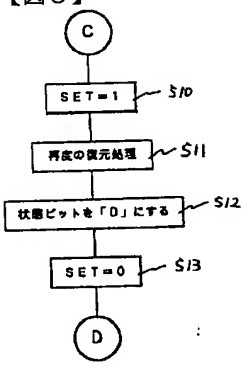
【符号の説明】

１００……サーバ計算機
１１０……ディスクサブシステム
１１１……ＳＣＳＩインタフェース
１１２……ＣＰＵ
１１３……ＲＯＭ
１１４……ＲＡＭ
１１５……バス
１１６……ＳＣＳＩインタフェース
１１７……ＳＣＳＩバス
１１８……磁気ディスク装置ＨＤ１
１１９……磁気ディスク装置ＨＤ２
１２０……磁気ディスク装置ＨＤ３
１２１……磁気ディスク装置ＨＤ４

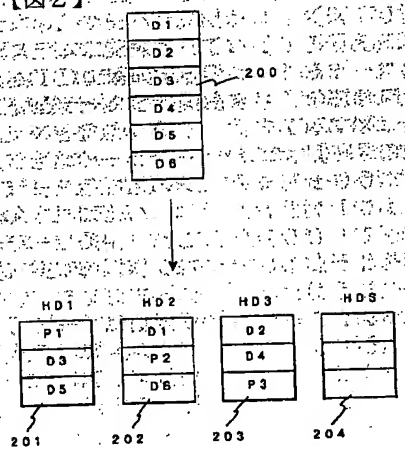
【図１】



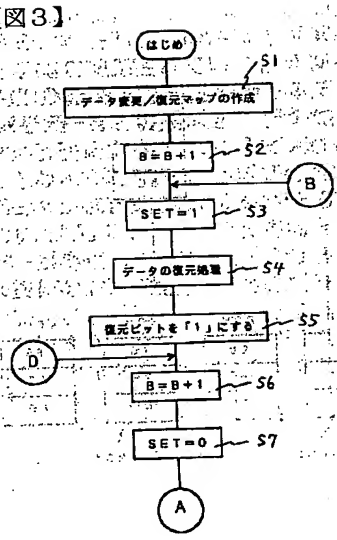
【図５】



【図２】



【図３】

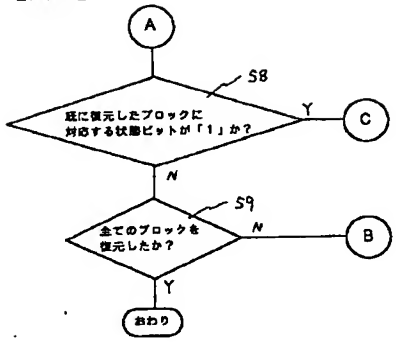


【図６】

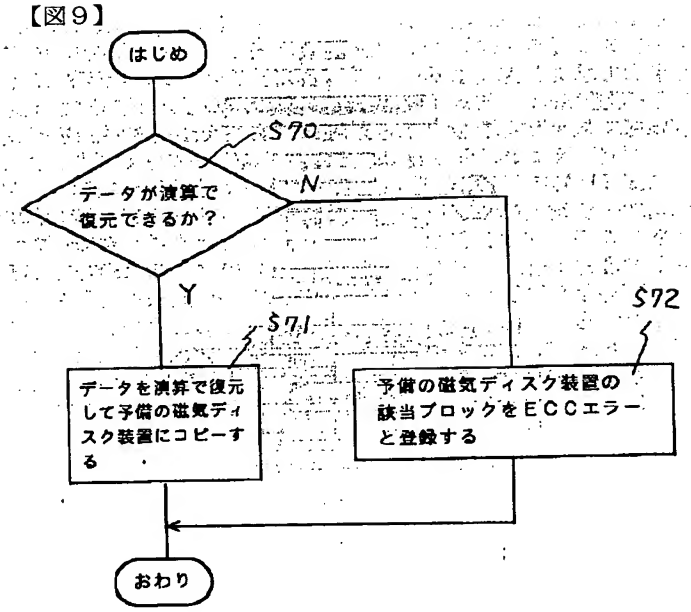
300

301	データ	D2	D4	P3
302	状態ビット	1	0	0
303	復元ビット	1	1	0

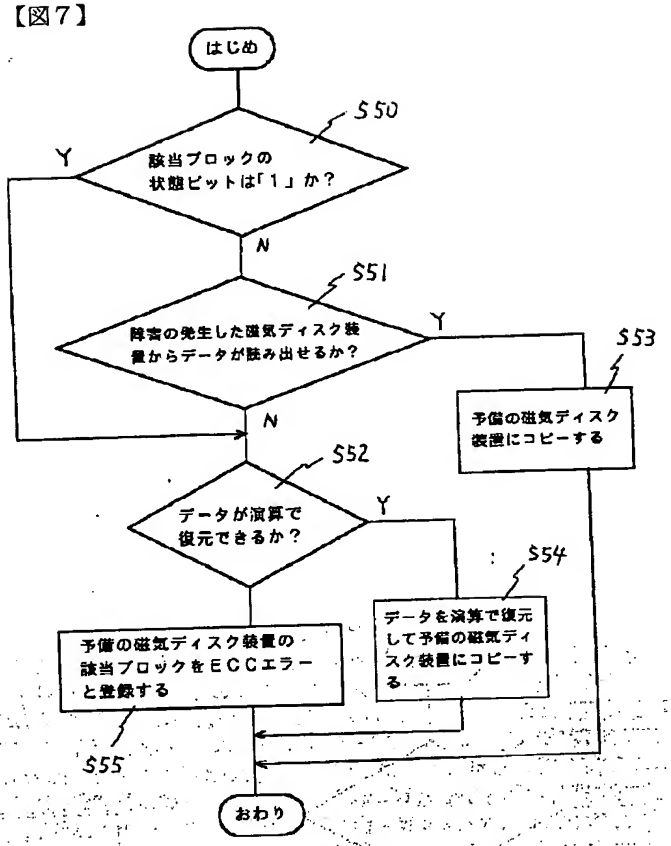
【図４】


A
S8
既に復元したブロックに
対応する状態ビットが「１」か？
Y
C
N
S9
全てのブロックを
復元したか？
N
B
Y
おわり

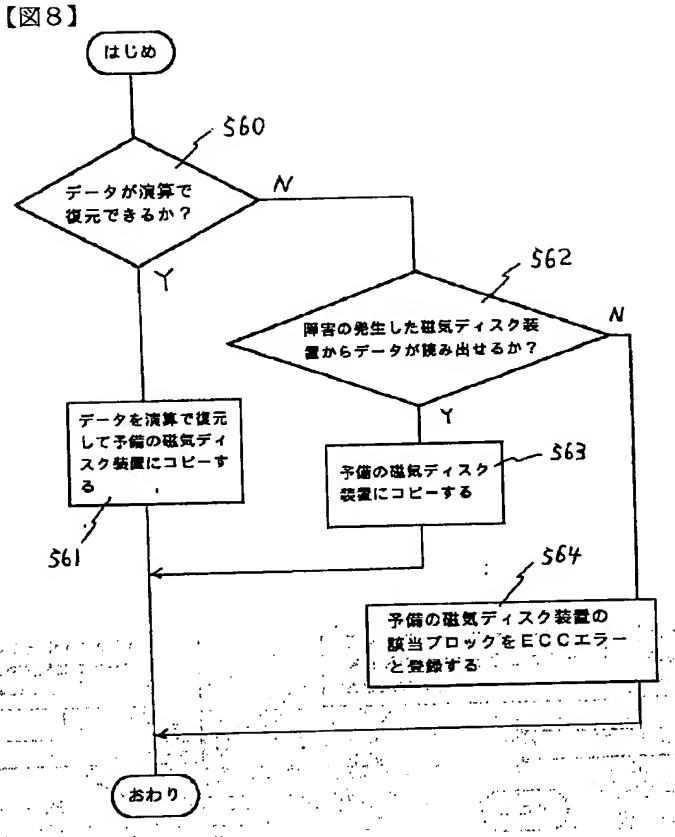
【図９】


はじめ
S70
データが演算で
復元できるか？
N
Y
S71
データを演算で復元
して予備の磁気ディ
スク装置にコピーす
る
S72
予備の磁気ディスク装置の
該当ブロックをＥＣＣエラー
と登録する
おわり

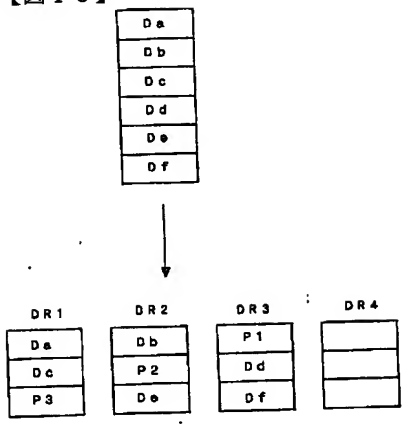
【図７】


はじめ
S50
該当ブロックの状態ビットは「１」か？
Y
N
S51
障害の発生した磁気ディスク装置からデータが読み出せるか？
Y
N
S53
予備の磁気ディスク装置にコピーする
S52
データが演算で復元できるか？
Y
S54
データを演算で復元して予備の磁気ディスク装置にコピーする
予備の磁気ディスク装置の該当ブロックをＥＣＣエラーと登録する
S55
おわり

【図８】


はじめ
S60
データが演算で復元できるか？
N
Y
S62
障害の発生した磁気ディスク装置からデータが読み出せるか？
N
Y
データを演算で復元して予備の磁気ディスク装置にコピーする
S61
予備の磁気ディスク装置にコピーする
S63
S64
予備の磁気ディスク装置の該当ブロックをＥＣＣエラーと登録する
おわり

【図１０】



フロントページの続き

(51)Int.Cl.7	識別記号	(51)F.I.	テーマコード（参考）
G06F 12/16	320	G06F 12/16	320L